Finding beans in burgers: Deep semantic-visual embedding with localization

Martin Engilberge[1,2], Louis Chevallier[2], Patrick Pérez[3], Matthieu Cord[1]
[1]Sorbonne Université, Paris, France [2]Technicolor, Cesson Sévigné, France, [3]Valeo.ai, Paris, France

Overview

Visual Grounding of phrases:

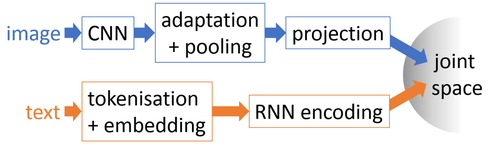
Localize any textual query into a given image.

Cross-modal retrieval:

Query: A cat on a sofa

Approach:

- Learning image and text joint embedding space.
- Visual grounding relying on the spatial-textual information modeling.
- Cross-modal retrieval leveraging the semantic space and the visual and textual alignment.



References

[1] K. He et al. Deep residual learning for image recognition. CVPR, 2016.
[2] T. Durand et al. Weldon: Weakly supervised learning of deep convolutional neural networks. CVPR, 2016.
[3] R. Kiros et al. Skip-thought vectors. NIPS, 2015.
[4] T. Lei et al. Training RNNs as fast as CNNs. arXiv, 2017.
[5] A. Eisenschtat et al. Linking image and text with 2-way nets. CVPR, 2017.
[6] F. Faghri et al. VSE++: Improved visual-semantic embeddings. arXiv, 2017.
[7] F. Xiao et al. Weakly-supervised visual grounding of phrases with linguistic structures. CVPR, 2017.

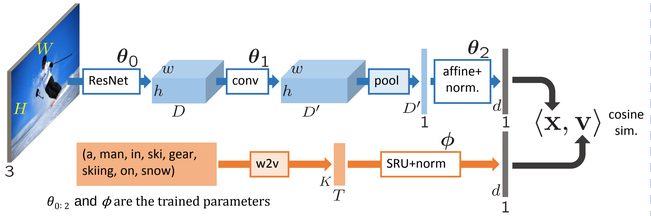
Semantic Embedding Model

Visual pipeline:

- ResNet-152 [1] pretrained.
- Weldon [2] spatial pooling.
- Affine projection and normalization.

Textual pipeline:

- Pretrained word embedding [3].
- Simple Recurrent Unit (SRU) [4].
- Normalization.



$\theta_{0:2}$ and ϕ are the trained parameters

Loss for a batch $\mathcal{B} = \{(\mathbf{I}_n, \mathbf{S}_n)\}_{n \in B}$ of image sentence pairs:

$$\mathcal{L}(\mathbf{\Theta}; \mathcal{B}) = \frac{1}{|B|} \sum_{n \in B} \left(\max_{m \in C_n \cap B} \text{loss}\,(\mathbf{x}_n, \mathbf{v}_n, \mathbf{v}_m) + \max_{m \in D_n \cap B} \text{loss}\,(\mathbf{v}_n, \mathbf{x}_n, \mathbf{x}_m) \right)$$

With :

- $\text{loss}(\mathbf{y}, \mathbf{z}, \mathbf{z}') = \max\{0, \alpha - <\mathbf{y}, \mathbf{z}> + <\mathbf{y}, \mathbf{z}'>\}$
- C_n (resp. D_n) set of indices of caption (resp. image) unrelated to n-th element.

Localization

Visual grounding module:

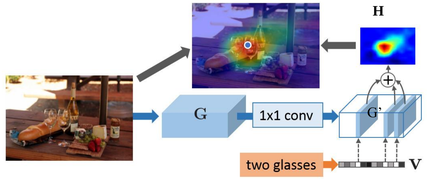
- Weakly supervised, with no additional training.
- Using the embedding space to select convolutionnal activation maps.

Generation of heatmap $\mathbf{H}$:

$$\mathbf{G}'[i, j, :] = A\mathbf{G}[i, j, :], \forall (i, j) \in [1, w] \times [1, h]$$

$$\mathbf{H} = \sum_{u \in K(\mathbf{v})} |\mathbf{v}[u]| * \mathbf{G}'[:, :, u]$$

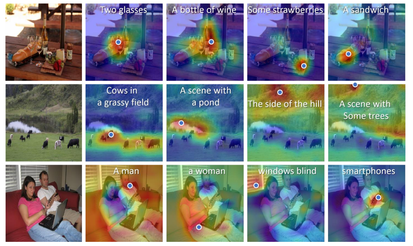
$K(\mathbf{v})$ the set of the indices of its k largest entries



Visual results

Visual grounding examples:

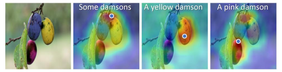
- Generating multiple heat maps with different textual queries.



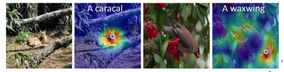
Toward zero-shot localization:

- Emergence of colors understanding:

- Even on artificial images:



- Generalization to unseen elements:



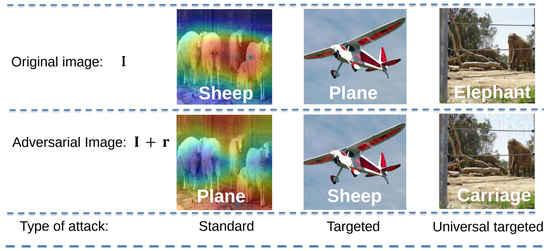
Sensitivity to adversarial attacks

Finding adversarial example:

Optimizing noise $\mathbf{r}$ over input image $\mathbf{I}$ resulting in its representation $F(\mathbf{I} + \mathbf{r}; \theta_{0:2})$ being displaced toward $\mathbf{y}$ in the embedding space.

$$\max_{\mathbf{r}} < F(\mathbf{I} + \mathbf{r}; \theta_{0:2}), \mathbf{y} > \text{ s.t } |\mathbf{r}|_2 \leq \boldsymbol{\varepsilon}$$

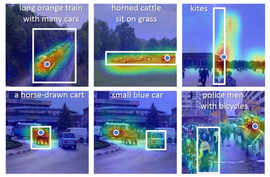
Examples of adversarial attacks



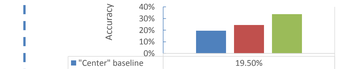
Closest word in the embedding overlaid in white.

Evaluation

The pointing game: Localizing phrases corresponding to subregions of the image.



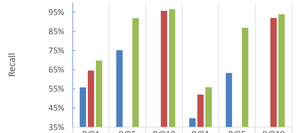
Pointing game results



"Center" baseline	19.50%
Linguistic structure [7]	24.40%
Ours	33.80%

Cross-modal retrieval: Evaluated on MS-CoCo image/caption pairs.

Cross-modal retrieval results

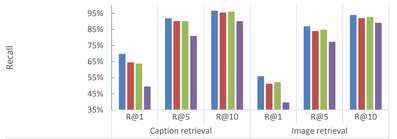


	Caption retrieval			Image retrieval		
2-Way Net [5]	55.80%	75.20%		39.70%	63.30%	
VSE++ [6]	64.60%		95.70%	52%		92%
Ours	69.80%	91.90%	96.60%	55.90%	86.90%	94%

Performance boost coming from:

- Architecture choice: SRU and Weldon spatial pooling.
- Efficient learning strategy: hard negative loss.

Ablation study: cross modal retrieval results



	Caption retrieval			Image retrieval		
Hard Neg + WLD + SRU 4	69.80%	91.90%	96.60%	55.90%	86.90%	94%
Hard Neg + GAP + SRU 4	64.50%	90.20%	95.50%	51.20%	84.00%	92.00%
Hard Neg + WLD + GRU 1	63.80%	90.20%	96%	52.20%	84.90%	92.60%
Classic + WLD + SRU 4	49.50%	81%	90.10%	39.60%	77.30%	89.10%

CVPR 2018 SALT LAKE CITY • JUNE 18-22